**Mini Paper**

**Music Enabled Running**

**“Can running be improved by using Music”**

**Mai Linh Luong**

[**m.luong@student.fontys.nl**](mailto:v.angelova@student.fontys.nl)

**Mitchel Kuijpers**

[**m.kuypers@student.fontys.nl**](mailto:m.kuypers@student.fontys.nl)

**Arwen**

[**a.voorn@student.fontys.nl**](mailto:a.voorn@student.fontys.nl)

**Filip Vangelov**

[**f.vangelov@student.fontys.nl**](mailto:f.vangelov@student.fontys.nl)

**Abstract** -

***I.INTRODUCTION***

Over the course of the Minor program in Applied Data Science, a team of 4 researchers was given the task to analyze multiple sets of data generated from foot sensors. The sensors were recording while the test runners were listening to music from Spotify. From there comes the question “Can running be improved by using Music”. This project covers extensive research on the different datasets generated by the sensors and other devices that were used in the data gathering.

***II.APPROACH***

The approach that was taken during this project is based on the IBM Data Science Methodology going through the phases of Business Understanding, Data Understanding, Data preparation, Modeling, Evaluation and Deployment.

***III.Data***

The datasets used during the music project consist of information gathered by two runners. The information was gathered with the help of footpods and the phones of the runners, which they took with them during each running session. The datasets contain data about the footpods, the music listened to by the runners, the music features, like energy and artist, phone locations, phone activity and the running sessions. During the project Olaf also delivered new data about gathered running sessions from the two runners, which was used to further help the progression of the project. The first two datasets created for the project consisted of the information from the footpods, sessions and music data. This was done because it was not yet necessary to use the phone data for that part of the project.

The two cleaned datasets for the first and second test runner were made with the help of multiple CSV files that were given by Olaf. These CSV files, as explained earlier contain data about the footpods, music and phone data. The dataset for the first test runner consists of 29 columns and 59457 rows and mostly contains object, float64, int64 and bool data types. The dataset for the second test runner consists of 20 columns and 6381 rows and also contains the same data types as the dataset for the first test runner.

After creating the cleaned datasets during the exploratory data analysis, new data was received to work with by Olaf. This dataset was already cleaned and contained multiple columns about the foot pod data. The newly received datasets contained new columns for the footpod features, which had their symmetry angle calculated. Like the previous datasets used during the EDA, these datasets also contained objects, float64, int 64 and bool data types. The dataset containing information about the first test runner contained 245,724 rows and 95 columns in total. The dataset containing information about the second test runner contained 17,840 rows and 95 columns in total.

***IV.GOALS***

The goal of this project is to find if there is a correlation between running performance and music.

***V.EXPERIMENT***

Because of the scale and goal of the project most of the focus was on cleaning the not prepared data sets that were initially generated from the sensors. For that the team made use of several methods for data cleaning and classification of clusters that might compromise the quality. Knn - means was used for creating clusters based on parameters such as (ex) contact\_time and impact. By applying the “Elbow rule” we determined the preferable number of clusters and removed the ones with the fewest data points. Frequency Column - To determine if the test person was running or walking, the frequency column was being added. Hence the calculation for frequency had to be done. The frequency shows how many steps the test person has taken in one second. The lower the frequency, the slower the test person was which means the more the test person was walking rather than running. Frequency Boxplot Outliers - with the help of the boxplot, outliers for frequency were removed. This way, only the running-rows of the dataset are being saved. Music Counts - before the test person was going for a run, he had to put on some sensors and turn on the app on the phone in order to save the running data. Every time one of the sensors or every time an activity is being done in the app, the application sees it as a new running session and therefore assigns for example a 2 minute test-run as a new session. This means that it was important to see which running sessions were valuable for the research. To filter out the wrong sessions, music count has been done. Hereby, when there is a session where less than 4 songs or more than 10 songs were played, it would’ve been considered a “bad” session. Bad sessions were being removed from the dataset. Duration Sessions - another way to filter out bad sessions was to filter out sessions with a short duration or sessions that were rather too long.

After the datasets were cleaned and new sets were provided the following actions were to find stable correlation between the data attributes generated from running and the music. For that the group made use of Correlation Analysis, Extreme Gradient Boosting, Self-Organizing Maps and developed a map covering the running session for better understanding (Kepler Map).

A correlation analysis has been done on both test subjects, as a goal to find running metrics correlated with music. Two used correlation methods are the Pearson and Spearman method. This concluded that the impact symmetry has the highest correlation with music for test subject 2, subject 1 didn’t have any correlation at all.

Extreme Gradient Boosting (XGB) was applied on the cleaned data in order to predict the impact symmetry. By applying the method a feature importance list was extracted, which gives insights what features have effect on the impact symmetry. This method was evaluated by using 5-fold cross validation, to decide the best parameters.

***VI.RESULTS***

The results of XGB for test subject 2; Mean Squared Error had a score of 0.196. Which is low, but taking in account that the range is between -0.242 & 0.173 while the mean is -0.049. Makes a score of 0.196 RMSE be less likely to be a positive thing. The Coefficient of determination (R2) score, is the proportion of the variation in the dependent variable that is predictable from the independent variable. Which is measured by a percentage, being close to 100% means a high correlation between the predictor and test. The score was -0.9%, this means that there is no relation between the xgb prediction and the impact symmetry score.

The results of the Self Organizing Map (SOM) didn’t lead to much success within the project even though there were interesting results. What happens in SOM is that each data point in the data set recognizes themselves by competing for representation. SOM mapping steps start from initializing the weight vectors. From there a sample is selected randomly and the map of weight vectors is searched to find which weight best represents that sample. Each weight bector has neighboring weights that are close to it.Therefore, the weights of the vectors didn’t tell much about the danceability of the running sessions of the test person.

***VII.CONCLUSION***

***AND RECOMMENDATIONS***

The project Music Enabled Running was quite unique compared to other projects. Instead of creating and having a clear goal, this project focused more on the scientific statement whether or not there was a correlation between the running performance and music.

During the project, it was clear that a lack of symmetry in the running performance often results in injury. One of the things that would be considered as bad running is when the impact symmetry is high. Stomping hard on the ground while running could also lead to serious injuries.

One of the things that could’ve been done when more time was reserved, is to try to have 2 test people included in this project. One person would run with music and the other person would run without music. Or let one test person run one session with music and one session without music. This way, a comparison can be made and eventually have a more clearer view whether or not there is an improvement in the running performance.

More data from test subject 2 could change the results of the correlation analysis. Comparing test subject 2 with 17,840 rows to test subject 1 with 245,724 rows, a case could be made that the correlations are because of the low data volume .

A strong statistical/mathematical understanding and evaluation is needed to decide the meaning/validity of the model output.